Wikipediaを用いた人名抽出と機械学習を用いたテレビ番組ジャンル推定

A genre estimation method of TV programs using the Wikipedia and a machine learning technique

福原 知宏[*1] 武田 英明[*1*2]
Tomohiro Fukuhara Hideaki Takeda

[*1]東京大学人工物工学研究センター
Research into Artifacts, Center for Engineering (RACE), The University of Tokyo

[*2]国立情報学研究所実証研究センター
Research Center for Testbeds and Prototyping, National Institute of Informatics

A genre estimation method for TV programs using the Wikipedia and a machine learning technique is described. The method uses a multiple-class support vector machine (SVM) for estimation of genres. Evaluation results using programs of a cable TV are reported.

1. はじめに

本論文では Wikipedia と機械学習を用いたテレビ番組ジャンル推定手法を提案する．今日，地上波放送，衛星放送に加え，ケーブルテレビ (CATV) やデジタル放送などにおいて様々な番組が製作され配信されている．これらの番組製作に伴い，番組の概要やジャンルといった番組メタデータの作成作業を効率化する必要がある．

本研究では番組メタデータのうち番組ジャンルに注目し，番組概要を用いたジャンル推定を行う．番組メタデータには番組名，放送チャネル，番組長さ，放送開始時刻，出演者名，数 10 字程度の概要文が含まれている．本研究ではこの番組メタデータを用い，日本語版 Wikipedia を用いて出演者名を抽出し，多クラス対応の support vector machine(SVM) を用いて ARIB 番組ジャンル 13 クラスへの分類を行った．提案手法の概要と正答率について述べる．

本論文の構成は次の通りである． 2. では先行研究について述べる． 3. では提案手法について述べる． 4. では実験結果について述べる． 5. では得られた結果について考察する． 6. で本論文の議論をまとめる．

2. 先行研究

テレビ番組のジャンル推定は基本的に機械学習における分類問題と見なすことができる．分類問題はこれまで様々な手法による解法が提案されてきた．この内，SVM を用いた分類器は様々な分野に応用されており，自然言語処理における応用 [工藤 04] や，スパムブログ (Splog) のフィルタリングへの応用例がある [Kolari 06]．一方，番組ジャンルは複数であることから，本研究では Crammer らの提案する多クラス対応の SVM[Crammer 02] を用いたジャンル推定を行う．

3. Wikipediaと機械学習を用いたテレビ番組ジャンル推定手法

本節では Wikipedia と機械学習を用いたテレビ番組ジャンル推定手法について述べる．本研究では以下のアプローチでジャンル推定を行う．

1. 番組データの収集
2. Wikipedia からの人名抽出
3. 番組ジャンルの学習と推定

以下，それぞれについて述べる．

3.1 番組データの収集

番組ジャンルの推定に当たり，番組データと番組ジャンル情報を収集した．番組データについては，ある CATV における 3 つのチャネルに含まれる番組と，地上波と衛星放送番組データを収集した．後者の収集に当たっては，Web 上のテレビ番組情報サイト[*1]のデータを用いた．CATV 番組も地上波・衛星放送番組も番組メタデータを持つが，CATV 番組についてはジャンルは付与されていない．表 1 に CATV の番組メタデータ例を示す．番組メタデータはタイトル，チャネル番号，放送開始時刻，番組長さ，概要からなる．地上波・衛星放送番組のメタデータには以下に述べる ARIB ジャンルデータが付与されている．

推定する番組ジャンルについては EPG(Electronic Program Guide) に付与されているジャンル体系を用いる．現在，地上波と衛星放送の番組ジャンルは ARIB(社団法人電波産業会) の規格に従っており，大項目として 13 ジャンルが存在し，更に下位のジャンルが各大項目に付与されている [社団 08]．本研究では ARIB 番組ジャンルの大項目 13 ジャンルの推定を行う．表 2 に大項目のジャンル一覧を示す．

表 3 に実験に用いる CATV の各チャネルの番組数 (再放送による重複を含む) と概要を示す．また，図 1 に各チャネルの 2008 年 2 月におけるジャンル構成を示す．いずれのチャネルも「情報／ワイドショー」と「ドラマ」を主体に構成されるが，Channel B は「アニメ／特撮」に，Channel C は「映画」にそれぞれ重点を置いていることが分かる．

3.2 Wikipedia からの人名抽出

本研究ではジャンル推定に際し，番組情報中のキーワードに加え，出演者名も特徴量として利用することから，日本語版 Wikipedia を用いた人名抽出を行った．Wikipedia には俳

連絡先: 福原知宏，東京大学人工物工学研究センター，千葉県柏市柏の葉 5-1-5，email:fukuhara@race.u-tokyo.ac.jp

*1 今回は日刊スポーツのテレビ番組欄 (http://www.nikkansports.com/) から番組情報を収集した．

表 1: 番組メタデータの例

タイトル	大河ドラマ　武田信玄#5
チャネル番号	999
放送開始時刻	18:00:00
番組長さ (分)	59
概要	＃５「湖水伝説」出演：中井貴一／平幹二朗／若尾文子／柴田恭兵／菅原文太　ほか　上洛を目前に天下取りの夢を果たせなかった戦国武将・武田信玄の生涯を描く。

表 2: ARIB による番組ジャンル区分

	ジャンル名 (1〜7)		ジャンル名 (8〜13)
1	ニュース／報道	8	アニメ／特撮
2	スポーツ	9	ドキュメンタリー／教養
3	情報／ワイドショー	10	劇場／公演
4	ドラマ	11	趣味／教育
5	音楽	12	福祉
6	バラエティー	13	その他
7	映画		

優や映画監督など多くの人名が登録されている．本研究では Wikipedia に記載されている人名を抽出し，人名抽出の精度向上を行った．

Wikipedia の記事データとして，2008 年 10 月 20 日の記事ダンプファイル[*2]を用いた．Wikipedia からのデータ抽出には Wikipedia データ解析ツール Wik-IE[*3]を用いた．抽出方法として，記事中に Inforbox が存在する場合は Inforbox 中から姓名を抽出し，Inforbox が存在しない場合は記事本文に対して正規表現を用いて姓名を抽出した．抽出した人名は MeCab のユーザ辞書に登録した．

以上の手続きを経て，最終的に Wikipedia から抽出した人名は計 20,493 人である．表 4 に抽出した人名カテゴリと人名数の一覧を示す．抽出元のカテゴリとして映画監督，俳優，落語家，声優などを用いた．

3.3　多クラス SVM によるジャンル推定

番組ジャンルが 13 ジャンル存在することから，本研究では多クラス SVM を用いてジャンル推定を行う．本論文では多クラス SVM の実装である $SVM^{multiclass}$[*4] version 2.20 を用いた．学習には線型カーネルを用い，正則化パラメータは $C = 5,000$ とした．

学習に用いる特徴として，(1) 番組長さ，(2) 出演者名，(3) 概要に出現するキーワードを用いた．出演者名については，番組メタデータ内の概要に対して CaboCha 0.53[*5][工藤 02] を適用し，人名抽出を行った．キーワードについては，番組概要に形態素解析を適用し，名詞，形容詞，副詞，未知語を抽出した．形態素解析器には MeCab 0.97[*6]を用いた．

*2　jawiki-latest-pages-articles.xml.bz2
*3　http://wik-ie.sourceforge.jp/
*4　http://svmlight.joachims.org/svm_multiclass.html
*5　http://chasen.org/~taku/software/cabocha/
*6　http://mecab.sourceforge.net/

表 3: CATV 各チャネルの概要 (番組数は再放送による重複を含む)

チャネル	番組数 (2 月分)	概要
Channel A	1,437	家族向けチャネル． ドラマ，スポーツ番組を多く含む．
Channel B	1,538	家族向けチャネル． アニメ／特撮番組を多く含む．
Channel C	932	家族向けチャネル． 映画を多く含む．

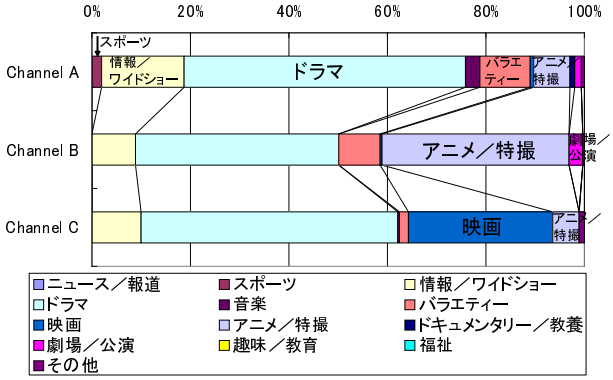

図 1: 実験に用いた CATV 各チャネルのジャンル構成

4.　実験結果

以下では実験結果について，(1)Wikipedia からの人名抽出による正答率の変化，(2)CATV 番組を訓練データとした場合の正答率，(3) 地上波と衛星放送番組表を用いた CATV 番組ジャンル推定の観点から述べる．

4.1　Wikipedia からの人名抽出による正答率の変化

Wikipedia を用いた人名抽出の効果を測定するため，人名抽出をしない場合と，した場合での正答率変化について調査した．ここでは CATV の Channel C を対象として実験を行った．訓練データには地上波と衛星放送の番組表 (2008 年 7 月 1 日〜10 月 31 日) から番組データ 30,000 件を用いた．テストデータには Channel C 2 月分のデータ 932 件を用いた．

MeCab のデフォルト辞書である mecab-ipadic-2.7.0-20070801 を用いた場合，正答率は 37.8% であった．これに対し Wikipedia による人名拡張を行った場合，正答率は 38.8% となり，大きな違いは見られなかった．大幅な改善は得られなかったものの，僅かな正答率改善は可能であることが分かった．今後，別のチャネル，別の時期のデータを使って効果を検証する．

4.2　CATV 番組を訓練データとした場合

CATV 番組に人手で正解ジャンルを付与し，訓練データとして用いた場合の正答率について述べる．

(1) 各チャネルの番組を訓練データとして推定した場合

各チャネルの番組を用いて予測した場合について述べる．訓練データとして，2008 年 2 月に放送された各チャネルの番組データを用いた．この訓練データを用いて，各チャネルの 2008 年 4 月分 (4 月 1 日〜4 月 30 日)，6 月分 (6 月 1 日〜30 日)，8 月分 (8 月 1 日〜31 日) の番組ジャンルを推定した．

表 4: Wikipedia から抽出した人名数

抽出元カテゴリ	抽出人名数
Category:日本の男性声優	1,455
Category:日本の女性声優	1,750
Category:落語家	853
Category:日本の俳優	7,038
Category:俳優に関するスタブ	6
Category:東京都出身の人物	8,692
Category:日本の映画監督	699
計	20,493

表 5: CATV 各チャネルの番組 (2 月分) を訓練データとした場合の正答率 (時期は 2008 年)

	訓練データ	テストデータ		
チャネル名	(2 月分)	4 月分	6 月分	8 月分
Channel A	99.9%	95.0%	96.2%	91.0%
Channel B	99.9%	95.5%	91.1%	85.3%
Channel C	99.5%	87.5%	88.7%	80.8%

表 5 に正答率を示す．結果，時期が下るにつれ正答率が下がっていることが分かった．これは毎月，各チャネルにおいて放送開始番組や放送終了番組があるためである．今後，時期が変わっても正答率を維持するための方策について検討する．

(2) 全 CATV チャネルの番組を用いて予測した場合

全 CATV チャネル (Channel A-C) の 2 月分の番組を訓練データとして用いて，他の月の番組ジャンルを推定した．表 6 に結果を示す．表 5 と比較して，Channel A 8 月分，Channel C 8 月分において正答率が多少改善されていることが分かる．一方，Channel B 8 月分の箇所を見ると，9.4 ポイント正答率が減少していることが分かった．現時点で原因は不明であるが，全チャネルの番組を混合したことによって正答率が低下する恐れがあることが分かった．

4.3 地上波＋衛星放送番組表からの CATV 番組ジャンル推定

4.2 では CATV の番組ジャンルを訓練データに用いることで，ある程度の正答率を得られることが分かった．一方，CATV 番組へのジャンル付与は手動であり作成コストが高いという問題がある．

ここでは既にジャンルが付与されている地上波と衛星放送の番組データを用いて，CATV 番組のジャンル推定を行った．訓練データには，2008 年 7 月 1 日〜10 月 31 日の期間に放送された地上波番組表 (東京地区) と衛星放送*7番組表を用いた．訓練データは計 30,000 件の番組情報 (再放送による重複を含む) を含む．テストデータには CATV の Channel A, B, C(2 月分) の番組データを用いた．訓練データ内での正答率は 77.3% であった．

表 7 に CATV 番組における正答率を示す．4.2 で見た CATV の番組ジャンルを用いた場合に比べ，全体的に正答率が低くなっている．正答率が高くなかった原因として，訓練に用いたデータ (地上波＋衛星放送番組表) と CATV の番組とで番組傾向が異なることが考えられる．すなわち，CATV では現在地上波で放送されていない過去の番組の再放送があったり，あるいは国内地上波では未放送の海外ドラマが放送されていることから，こうした場合に現在の訓練データでは対応できなかったと考えられる．特に Channel B は 20% 台の正答率であり，地上波と衛星放送の番組だけでは推定が難しいことが分かる．

*7 NHK BS1，NHK BS2，WOWOW，NHK ハイビジョン，BS 日テレ，BS 朝日，BS-TBS，BS ジャパン，BS フジ，BS イレブン，トゥエルビ，スター・チャンネル BS を含む．

表 6: CATV 全チャネルの番組 (Channel A から C の 2 月分) を訓練データとした場合の正答率 (時期は 2008 年)

	訓練データ	テストデータ		
チャネル名	(2 月分)	4 月分	6 月分	8 月分
Channel A	99.6%	94.9%	96.0%	94.8%
Channel B	99.9%	94.2%	96.0%	75.9%
Channel C	99.4%	91.8%	88.2%	83.7%

表 7: 地上波＋衛星放送番組情報を用いた CATV 番組ジャンル正答率 (テストデータは各チャネルの番組 2 月分)

チャネル	正答率 (%)
Channel A	59.9
Channel B	29.7
Channel C	40.8

図 2 に各チャネルのジャンル別 F 値を示す．ドラマについてはいずれのチャネルとも良好な F 値となっている一方で，Channel B では情報／ワイドショーの F 値が低かったり，バラエティーについては各チャネルとも F 値が低いことなどから，ジャンル毎の F 値の改善策について検討する必要がある．

4.4 正解データのフィードバックによる CATV 番組ジャンル推定

4.3 で見たように，地上波＋衛星放送番組表だけでは十分な正答率が得られなかった．ここでは地上波＋衛星放送番組表に加え，CATV の正解番組ジャンルを訓練データに混ぜた場合の正答率について述べる．

訓練データには 2008 年 7 月 1 日〜10 月 31 日 (123 日) の地上波番組表 (東京地区) と衛星放送番組，計 30,000 件を用い，これに加え，CATV 各チャネルの 2 月分番組データ (正解ジャンル付与済み) を用いた．CATV の正解データの投入量として 1 日分，2 日分，3 日分，7 日分，14 日分，全正解データ (1ヶ月分) の組み合わせについて調査した．テストデータには CATV 各チャネルの 2 月分データを用いた．表 8 に結果を示す．

結果，1 日分の CATV 正解データをフィードバックするだけでも大幅な正答率改善が得られることが分かった．ここでは訓練データとテストデータに同じ月のデータを用いたが，今後，CATV の他の月の番組を用いて評価を行う．

5. 考察

本節では，実験結果についての考察と今後の課題について述べる．

5.1 実験結果についての考察

1. ドラマと映画の区別
2. 番組製作時期による出演者役割の違い
3. 複数のジャンルに登場する出演者の考慮
4. 海外俳優の取り扱い

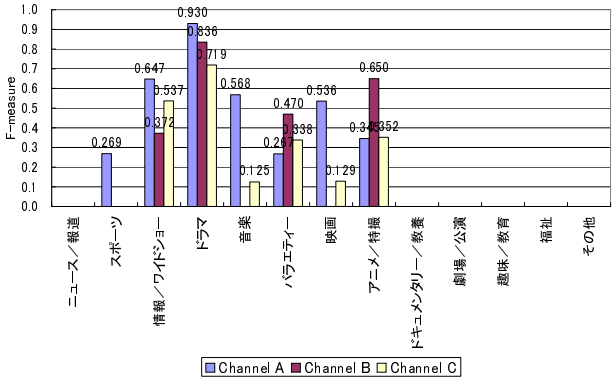

図 2: CATV 各チャネルのジャンル別の F 値

表 8: CATV 番組正解データのフィードバックによる正答率変化 (括弧内はベースラインからの増加ポイント数)

フィードバック量	Channel A	Channel B	Channel C
ベースライン	59.9%	29.7%	40.8%
1 日分	79.2%(+19.3)	72.4%(+42.7)	67.2%(+26.4)
2 日分	79.4%(+19.5)	82.3%(+52.5)	67.8%(+27.0)
3 日分	81.1%(+21.2)	88.4%(+58.7)	64.6%(+23.8)
7 日分	81.4%(+21.4)	96.7%(+67.0)	63.2%(+22.4)
14 日分	83.2%(+23.2)	95.8%(+66.1)	79.1%(+38.3)
全データ	93.7%(+33.8)	98.5%(+68.8)	88.8%(+48.1)

第 1 に，ドラマと映画の区別の問題がある．ドラマによっては番組長さも映画並みのもの*8もあり，こうした場合に推定ジャンルを誤るケースが見られた．映画とみなすことのできるドラマも存在するが，今後，特徴量の選択などによって正しく推定できるよう検討する必要がある．

第 2 に，番組製作時期による出演者役割の違いがある．地上波をはじめ CATV では過去に製作された番組を再放送している場合が多く見られる．この場合，その番組の出演者の役割が時を経て変わる場合があり*9，訓練時のジャンルと結びつかないという問題がある．今回，地上波と衛星放送の番組を用いて CATV 番組のジャンル推定を行ったが，CATV では地上波・衛星放送で放送されていない過去の番組を放送している場合もあるため，この場合の正答率の改善について検討する必要がある．

第 3 に，複数のジャンルの番組に登場する出演者について検討する必要がある．例えば，ビートたけしはコメディアン，俳優，映画監督として活躍しているが，番組の製作時期によって番組ジャンルは異なる．今回，ビートたけしの出演する「風雲！たけし城」がバラエティではなくドラマに分類されていた．これは訓練データに出演者の過去の出演番組の情報がないためと考えられる．この場合の対策についても検討する．

第 4 に，海外俳優の取り扱いも検討事項である．CATV を中心に海外ドラマが放送されているが，それらドラマの出演者に関する情報が少なく，予測精度を改善できない．Wikipedia を用いて海外俳優に関する情報を収集するなどの対策が必要である．

*8 例えばドラマスペシャル，新春大型時代劇など．
*9 例えば明石家さんまは俳優としてドラマに出演していた時期があるが，現在はバラエティー番組を中心に出演している．

5.2 今後の課題

今後の課題として以下を挙げる．

1. 共演者情報の利用
2. Wikipedia を用いた出演者属性の利用
3. 他のカーネル関数の選択
4. 正解データのフィードバックによる再学習

第 1 に，共演者情報の利用について検討する．あるジャンルの番組に出演する出演者は，他の番組でも共に出演する場合がある．この時，出演者間の共出演関係から共出演関係の高い出演者を追加することで予測精度の向上が可能ではないかと筆者らは考える．このため，今後，同じ番組に出演したことのある出演者のネットワーク (出演者ネットワーク) を用いて，共演者を追加した場合の推定精度について調査を行う．

第 2 に，Wikipedia を用いた出演者属性を特徴量として取り込むことを検討する．Wikipedia には，俳優の生年月日やカテゴリ (例: アクション俳優)，出演作品などが記載されている場合が多い．こうした情報も特徴量として取り込むことで，ジャンル推定精度の向上を期待できる．今後，Wikipedia に記載されている出演者の関連情報の利用について検討する．

第 3 に，他のカーネル関数を用いた場合についても検討する．本論文では線型カーネルだけを用いたが，他のカーネル関数を用いても実験を行い，カーネル関数による推定精度の違いについて検討する．

第 4 に，誤った推定結果に対して判定者が手動で修正したデータを分類器にフィードバックした場合の分類性能について調査を行う必要がある．提案手法を実サービスとして実装する場合，強化学習のようにフィードバックループを持つシステムとして設計した方が良いと筆者らは考える．本論文でも CATV の正解データをフィードバックした場合の正答率について調査を行ったが，今後，他のチャネル，他の時期のデータについても調査する必要がある．

6. まとめ

本論文では Wikipedia と機械学習を用いた TV 番組ジャンル推定手法を提案した．地上波放送と衛星放送番組のデータ，CATV 番組データに多クラス SVM を適用し，CATV 番組のジャンル推定を行った．CATV の正解データをフィードバックすることで正答率を改善できることを確認した．今後の課題として，共演者情報の利用，Wikipedia からの出演者属性情報の利用による正答率改善について取り組む．

参考文献

[Crammer 02] Crammer, K. and Singer, Y.: On the algorithmic implementation of multiclass kernel-based vector machines, *J. Mach. Learn. Res.*, Vol. 2, pp. 265–292 (2002)

[Kolari 06] Kolari, P., Java, A., Finin, T., Oates, T., and Joshi, A.: Detecting Spam Blogs: A Machine Learning Approach, in *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI 2006)* (2006)

[工藤 02] 工藤 拓, 松本 裕治：チャンキングの段階適用による日本語係り受け解析, Vol. 43, No. 6, pp. 1834–1842 (2002)

[工藤 04] 工藤 拓, 松本 裕治：カーネル法を用いた言語解析における高速化手法, 情報処理学会論文誌, Vol. 45, No. 9, pp. 2177–2185 (2004)

[社団 08] 社団法人電波産業会：デジタル放送に使用する番組配列情報 (ARIB STD-B10), 4.6 版 (2008), (http://www.arib.or.jp/english/html/overview/doc/2-STD-B10v4_6.pdf)